DOCUMENT RESUME

ED 441 867 TM 031 209

AUTHOR	O'Leary, Michael
TITLE	The Effects of Age-Based and Grade-Based Sampling on the Relative Standing of Countries in International Comparative Studies of Student Achievement.
PUB DATE	2000-04-00
NOTE	32p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Opinion Papers (120) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Academic Achievement; Achievement Tests; *Adolescents; *Age Differences; Comparative Analysis; Foreign Countries; *Instructional Program Divisions; International Education; International Studies; Junior High Schools; *Science Achievement; *Science Tests; Sciences; *Test Results
IDENTIFIERS	*Ireland; Third International Mathematics and Science Study

ABSTRACT

The investigation reported in this paper was prompted by discrepancies between the published outcomes from two international tests of science achievement: (1) the Second International Assessment of Educational Progress (IAEP2), administered in 1991; and (2) the Third International Mathematics and Science Study (TIMSS), administered in 1995. One finding was that while average science achievement for Irish 13-year-olds was reported to be at the low end of the distribution for the 20 participating countries in IAEP2, it was around the middle of the distribution for the 40 or so countries that participated in TIMSS in the early grades of secondary schooling. Initial comparisons suggested that there were also inconsistencies in outcomes for some of the 11 other countries that participated in both surveys, such as France, Portugal, and Switzerland. Analyses reveal that when sampling/population definition differences between the two surveys are accounted for, science achievement in Ireland was not at the low level suggested by initial interpretations of IAEP2 data but was closer to the levels reported in TIMSS. While the sampling issue did not fully account for discrepancies with respect to the IAEP2/TIMSS outcomes for some countries, it is argued that the findings outlined in this paper have a number of implications for policymakers using data from future international comparative studies of student achievement. (Contains 2 figures, 6 tables, and 23 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 867

The Effects of Age-Based and Grade-Based Sampling on the Relative Standing of Countries in International Comparative Studies of Student Achievement

Michael O'Leary
St Patrick's College
Dublin
Ireland

Paper presented at the annual meeting of the American Educational Research Association
New Orleans
April 24-28

TM031209

M. O'Leary

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

The Effects of Age-Based and Grade-Based Sampling on the Relative Standing of Countries in International Comparative Studies of Student Achievement

Abstract

The investigation reported in this paper was prompted by discrepancies between the published outcomes from two international tests of science achievement: the Second International Assessment of Educational Progress (IAEP2) administered in 1991and the Third International Mathematics and Science Study (TIMSS) administered in 1995. One finding was that while average science achievement for Irish 13-year-olds was reported to be at the low end of the distribution for the 20 participating countries in IAEP2, it was around the middle of the distribution for the 40 or so countries that participated in TIMSS in the early grades of secondary schooling. Initial comparisons suggested that there were also inconsistencies in outcomes for some of the 11 other countries that participated in both surveys e.g. France, Portugal, and Switzerland. Analyses described here reveal that when sampling/population definition differences between the two surveys are accounted for, science achievement in Ireland was not at the low level suggested by initial interpretations of the IAEP2 data but was closer to the levels reported in TIMSS. While the sampling issue did not fully account for discrepancies with respect to the IAEP2/TIMSS outcomes for some countries, it is argued that the findings outlined in this paper have a number of implications for policy makers using data from future international comparative studies of student achievement.

The Effects of Age-Based and Grade-Based Sampling on the Relative Standing of Countries in International Comparative Studies of Student Achievement

In international comparative studies of student achievement, a number of countries (usually represented by research organisations) agree on an instrument to assess achievement in a curriculum area, the instrument is administered to a representative sample of students at a particular age or grade level in each country, and comparative analyses of the data obtained are carried out. The potential of such studies to contribute to policy formation in many areas was made clear from the earliest studies in the 1960s and in subsequent years. The areas include the pursuit of equity goals, setting priorities, assessing the effectiveness and efficiency of the educational enterprise and the appropriateness of curricula, evaluating instructional methods and the organisation of school systems, and providing a mechanism for accountability (Kellaghan & Grisay, 1995; Plomp, 1992). While there is relatively little information on the extent to which the findings of studies have in fact been utilised for any of these purposes, there is no doubt that they attract considerable media and public attention.

Goldstein (1997) points out that the preferred ages for testing in international comparative surveys are 9- and 10-year-olds, 13- and 14-year-olds and those students in the final year of secondary school. When an age-based population definition is used, pupils of a particular age (e.g. 13-year-olds) are sampled and tested. When the focus is on grade, pupils in a particular grade (e.g. 7^{th} grade) are sampled and tested. When students are sampled by age the intention is that the maturity level of students are as similar as possible. Moreover, an age-based sampling approach is efficient in so far as it produces more reliable student-level estimates by minimizing the effects of clustering within

schools and also requires smaller sample sizes (Foy, 1998). However, sampling by age often results in dissimilar educational experiences in so far as students of a similar age may be in two or three different grades. Jaeger (1994), for example, has pointed out that at given ages, students in Korea and Taiwan have received substantially more schooling than their counterparts in the US. It can be shown that students in countries such as Ireland, who start school at five years of age, will have eight years of formal schooling by age thirteen. Students who start school at six or seven will have less. For these reasons, sampling by age also has the associated disadvantages of making it more difficult to construct different "causal" models involving pedagogy and curriculum, to align tests to curriculum and to administer a test to students in different classrooms (Foy, 1998). Contrariwise, if sampling is done by grade, then the amount of formal schooling will usually be similar but the age of students in any particular grade may differ by as much as a year. The average age of eight graders in TIMSS ranged from 13.6 years in Iceland to 14.6 years in Iran. The range was even bigger when all countries, not just those satisfying sampling guidelines were considered (see Beaton et al., 1996, p. 22). In addition, comparability when grade sampling is employed is complicated by the need to account for policies related to grade repetition and promotion within school systems (Goldstein, 1995, 1997).

Wiley and Wolfe (1992) argue that since grade definitions may vary between countries, an age-based population definition may allow for better international comparability. However, it is also acknowledged by many that strict comparability in international assessments may not be possible (e.g., Keeves, 1992c) no matter what design is utilised. In TIMSS, comparability problems with respect to age and grade sampling were addressed by administering the same test to the pair of adjacent grades containing

most of the students of interest (9- and 13-year olds). The planners felt that this approach facilitated analyses where homogeneity of age, homogeneity of curricular experiences or both were important considerations (Beaton, Martin, & Mullis, 1997).

The purpose of the study described in this paper is to examine how the use of age- or grade-based population definitions can affect the relative standing of some countries in international comparisons of achievement. The paper begins with a brief outline of the Second International Assessment of Educational Progress in Mathematics and Science [IAEP2] (Lapointe, Askew, & Meade, 1992) and the IEA's Third International Mathematics and Science Study [TIMSS] (Beaton, Martin, Mullis, Gonzalez, Smith, & Kelly, 1996a). An important difference between the two studies is that while sampling was age-based in IAEP2, it was grade-based in TIMSS. In the second section of the paper, the focus switches to the science performance of 12 countries that participated in both studies at the early grades of secondary school. First, the published results are reviewed and particular attention is paid to the apparent discrepancies in the performance of a number of countries, including Ireland, across the two studies. Then, outcomes for students matched for age and grade are examined and compared and contrasted with the published results. The paper concludes with a discussion about the importance of considering population definition differences when evaluating the outcomes of international comparative assessments.

An Overview of IAEP2 and TIMMS

A total of 20 countries participated in IAEP2 though not all countries had comprehensive populations (see Lapointe, Askew, & Meade, 1992). Representative samples of 9-year-olds (born in 1981) and 13-year-olds (born in 1977) were tested in

mathematics and science. A number of countries also participated in a geography assessment and in a mathematics and science performance assessment. The IAEP2 science test for 13-year olds was contained in a single booklet, which had to be completed by students in four 15-minute segments (one hour of testing time in all). The science test consisted of 72 items and covered four content areas: Earth/Space Sciences, Life Sciences, Physical Sciences, and the Nature of Science. Eight items were excluded from the final analysis of student achievement due to the fact that they exhibited high differential item functioning (DIF).[1] Six of these items came from the Life Science content area. The other two came from the content areas of the Physical Sciences and the Nature of Science (Educational Testing Service [ETS], 1992a).

In all, 45 countries participated in TIMSS. However, a number of countries did not satisfy guidelines for sample participation rates, had unapproved sampling procedures, or had unapproved age/grade specifications (see, for example, Beaton et al., 1996). In each country, TIMSS tested the mathematics and science achievements of students in the grades containing most 9-year-olds (equivalent to 3rd and 4th grades in most countries), most 13-year-olds (equivalent to 7th and 8th grades in most countries) and in the final year of secondary education. Unlike the IAEP2 design, the TIMSS test booklets contained both mathematics and science items. At the seventh and eight grades the mathematics test was comprised of 151 items and the science test was comprised of 135 items. All items were rotated across eight test booklets and student performance on these booklets were matrix sampled using a modified Balanced-Incomplete-Block spiraling (BIB) design (Beaton, Martin, & Mullis, 1997). Each booklet was completed by students in two timed blocks of 44 and 46 minutes -- a total of one and one half hours of testing time in all. Together the TIMSS science items covered five

[1] Dorans & Holland (1993) and Holland & Thayer (1988) provide a good overview of DIF.

content areas: Chemistry, Earth Science, Environmental Issues/Nature of Science, Life Science, and Physics.

A total of 12 countries participated in both the IAEP2 and TIMSS studies of achievement at the early grades of secondary schooling. These countries were Canada, England, France, Hungary, Ireland, Korea, Portugal, Scotland, Slovenia, Spain, Switzerland, and the US. While Israel participated fully in IAEP2, it participated only at the eighth grade in TIMSS. In addition, its sampling procedures at classroom level were unapproved and it also failed to meet other study guidelines. Hence, Israel will be excluded from the analysis here. For ease of reference the twelve countries will be referred to as the IAEP2/TIMSS "common" countries or "common" countries for short.

A Comparison of the IAEP2 and TIMSS Published Results

Table 1 presents the IAEP2 and TIMSS average scale scores in science for the 12 countries that participated in both surveys of achievement at the early grades of secondary schooling. Countries are listed from highest achieving to lowest achieving and are categorised according to whether their means were statistically significantly above, below or not significantly different to the Irish mean.

[Insert Table 1 about here]

Relative to other countries, the performance of Irish students in IAEP2 science was very poor. As is shown in Table 1, Irish student performance compared unfavorably with performance in most other common countries. In terms of statistical significance the average achievement of students in Ireland was lower than in all comparison countries except Portugal and the US. The average proficiency score in Ireland was also

significantly below the IAEP2 overall average for all 20 participating populations (Martin, Hickey, & Murchan, 1992). In Ireland it was reported that that the average science achievement of Irish students in IAEP2 was not much better than the achievement of Korean and Swiss students at the tenth percentile. It was also pointed out that that even Irish students at the 90th percentile were not substantially better than the average achieving students in both of these countries. One had to search around the Irish 97th or 98th percentiles to find students that compared favorably with the top 10% of students in most other countries (Martin, Hickey, & Murchan, 1992). The poor performance of Irish students in science was also a feature of the first International Assessment of Educational Progress (IAEP1) (Lapointe, Mead, & Phillips, 1989).

In TIMSS, Irish students performed significantly above the average for all participating countries at both grade levels (Beaton et al., 1996). At the seventh grade, Ireland's performance was on a par with such countries as Canada, Switzerland, and the US, but was significantly better than the average performance in France, Portugal, Scotland and Spain. At the eighth grade, the Irish average was (statistically) significantly above the Swiss average. Two countries, Korea and Slovenia, achieved averages that were significantly higher than Ireland's at both grade levels. The comparison with the Swiss is particularly significant in so far as Swiss science performance in IAEP2 was so clearly superior at every level (low, average, and high achieving students) to the Irish performance. Such comparisons give the impression that, *relative to other countries*, Ireland's performance improved considerably from IAEP2 to TIMSS.

Clearly, the TIMSS findings for Ireland were surprising in so far as the pattern of poor Irish performance in science set in IAEP1 and IAEP2 was not continued.[2] It should also be apparent that the relative science performances of countries other than Ireland also seemed to change between IAEP2 and TIMSS. For example, the relative position of French and Swiss students in the two assessments differed markedly. It is not difficult to imagine the dilemmas posed by these findings for policy makers and others. In Ireland, the possibility that the findings (for some countries at any rate) indicated a change in the level of science achievement over time was considered but was rejected due to the fact that the studies were considered just four years apart. While many other hypotheses were considered (see O'Leary, 1999; O'Leary, Madaus & Kellaghan, 1997), the investigation of the effects of population definition differences in the two studies provided important clues about why findings in the two studies appeared anomalous.

The Distribution of Sampled 13-year-olds in IAEP2 Across Grades.

Table 2 contains the weighted percentage of 13-year-olds that were in grades 7 and 8 when they took the IAEP2 test in 1991. Also included in the table is the weighted percentage of 13-year-olds who were in grades outside the two most common grades.

[Insert Table 2 about here]

[2]TIMSS science results for students in the primary school were also surprising (see Martin, Mullis, Beaton, Gonzalez, Smith, & Kelly, 1997). At both grade levels (3rd and 4th class) Irish students did much better than might have been expected given the poor performance of Irish 9-year-olds in IAEP1 (see Lapointe, Meade, & Phillips, 1988).

It should be evident from Table 2 that in IAEP2 while most of the students tested in Ireland and Slovenia were in the grade 7, the majority of students in the other ten countries were in grade 8. The situation in Scotland is particularly revealing. Here, only 0.5% of the 13-year-olds tested in IAEP2 were at the grade where most Irish 13-year-olds were (grade 7). In fact, 86% of the Scottish 13-year-olds were in grade 8 and a further 13.5% were beyond that. In other words, almost all Scottish students tested in IAEP2 were one year further along in their secondary schooling than were the majority of the Irish test takers. The same is true for almost three-quarters Swiss students. Indeed, it is significant that, Slovenia apart, Ireland had the smallest percentage of 13-year-olds in grade 8 in IAEP2. And since students further along in their schooling achieve higher average scores (see Tables), it seems reasonable to argue that this was at least part of the reason why countries such as Canada and Spain (with 79.9 and 79.0% of students in grade 8 respectively) outperformed Ireland in IAEP2 when outcomes for 13-year-olds regardless of grade were presented. The extent to which this argument is supported by the achievement data is now considered in the following section.

A Comparison of IAEP2 and TIMSS Outcomes for Students Matched by Age and Grade

A crucial difference between the design of the IAEP2 and TIMSS surveys was that the former sampled by age while the latter sampled by grade. In addition, while the grades containing most 13-year-olds were sampled in TIMSS, the definition of age used in the two surveys was different (see Lapointe, Askew, & Mead, 1992; Martin & Kelly, 1996). In IAEP2, age was defined by calendar year of birth - students born in 1977 and taking the test in the Spring of 1991 were defined as 13-year-olds. This definition meant that students born in the first months of 1977 were actually 14 years old when they took the test. In

TIMSS, the 13-year-old cohort was determined with reference to the time of testing – students had to be 13-years old when tested. In other words, the IAEP2 definition of 13-year-olds resulted in students being, on average, up to four months older than 13-years-old as defined in TIMSS. As a result, while scale scores for 13-year-olds were included in the TIMSS reports (see Beaton et al., 1996, p. 37), they were not directly comparable with the IAEP2 results.[3] Therefore, samples of students surveyed in IAEP2 and TIMSS were matched for age and grade. First, IAEP2 13-year-olds (i.e. born in 1977) that were in grade 7 or grade 8 (as defined in TIMSS) were identified. Second, as TIMSS testing took place in 1995, students born in 1981 (i.e., 13-year-olds as defined in IAEP2) that were in grade 7 or grade 8 were identified. Outcomes for the matched samples of students are included in Table 3.

An important issue to bear in mind when interpreting the results to follow is that the proportions of 13-year-olds at a grade level in certain countries was small. This is especially true for the IAEP2 survey where far fewer students were tested than in TIMSS. Clearly, the fact that there was a very small proportion of Scottish 13-year-olds in IAEP2 (0.5%) at the seventh grade level rules out the possibility of making useful comparisons with TIMSS in this case.[4]

[Insert Table 3 about here]

[3] In addition, as TIMSS sampled by grade a median rather than an average scale score was estimated.
[4] The proportion of Scottish 13-year-olds (born in 1981) in the seventh grade in TIMSS was also small (3.1%).

The top and bottom halves of Table 3 contain average scale scores, standard errors and standard deviations for 13-year-old students in grades 7 and 8 respectively. The left-hand side of the table contains the IAEP2 results, the right-hand side the TIMSS results.

In terms of rankings and significant differences, the results for the seventh grade presented in Table 3 seem somewhat more consistent than the results discussed earlier. They show that Irish average performance was not significantly different from Canadian and US performance on either testing occasion. In addition, Korea, Hungary and Slovenia ranked higher and Spain and Portugal ranked lower on both occasions. English/Irish comparisons also seem consistent despite the fact that in IAEP2 there was no statistically significant difference between the average scores at the seventh grade (due principally to the large standard error associated with the English average). These results are encouraging as it will be recalled from the earlier analysis (see discussion around Table 1) that in four of these cases (Canada, Spain, Portugal and the US) comparisons in Ireland across IAEP2 and TIMSS seemed problematic. However, that said, the comparisons between outcomes for Irish students and their French and Swiss counterparts still seem anomalous at the seventh grade level. Even when the age and grade of students is accounted for, it can be seen that the rankings change. While Ireland and France achieved almost identical averages in IAEP2, Ireland's average in TIMSS was significantly higher. The comparison of Irish and Swiss averages at the seventh grade is also problematic. In IAEP2, Swiss 13-year-olds achieved significantly better averages than their Irish counterparts. In TIMSS the comparison was reversed at the seventh grade.

At the eighth grade level, the problem with the Irish/French and Irish/Swiss rankings persists. There is also the added inconsistency of the Irish/Portuguese and Irish/Scottish rankings across the two surveys. Indeed it could be argued that, in

comparison with the grade 7 results, the consistency of the grade 8 results across IAEP2 and TIMSS is not improved greatly when performance is broken out by age and grade.

Another issue of note raised by these data pertains to differences across countries with respect to the influence of grade on average performance. Figure 1 was constructed using data in Tables 1 and 3 to illustrate the differential between grade level averages in both surveys. The first column pertains to grade level differences in average achievement in IAEP2. The second column shows differences for 13-year-olds (born 1981) in TIMSS. The third column shows average differences between grade averages for the whole TIMSS cohort (i.e. as contained in the published reports) Differences between grade level averages in a country are expressed in terms of an effect size.[5]

[Insert Figure 1 about here]

---

[5] The effect size is a measure of the magnitude in numerical terms of a difference of interest (in the present case, mean differences between countries) (Hair, Anderson, & Black, 1995; Wolf, 1986). Its calculation involves dividing the value of the difference between two group means by the pooled standard deviation, a procedure which provides a scale-invariant estimate of the magnitude of the effect. This is accomplished using:

$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$ where

d is the effect size index for differences between means in standard units;

$\bar{x}_1$ and $\bar{x}_2$ are the sample means in original measurement units; and

s_{pooled} is the pooled standard deviation for both samples and is calculated as

$$\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

The effect size measure is now in the common metric of standard deviation units. Thus, an effect size of 0.3 indicates that one country scored 0.3 of a standard deviation higher (or lower) than the comparison country. In the literature, guidance for interpreting effect sizes is equivocal. Cohen (1977) has interpreted effect sizes around 0.2 as small, those around 0.5 as medium, and those around or above 0.8 as

BEST COPY AVAILABLE

On average, countries whose 13-year-olds were in the eighth grade when they took the IAEP2 test scored about 0.63 of a standard deviation unit above their counterparts in the seventh grade. Interestingly, the exact same difference was found to apply in the case of TIMSS 13-year-olds. The difference between grade averages in TIMSS when all students were considered was lower and amounted to an effect size of 0.44. In IAEP2, the Irish average at the eighth grade level was exactly half of a standard deviation higher (effect size = 0.50) than at the seventh grade level. A similar difference was found to occur in TIMSS whether one considers only 13–year-olds in the two grades (effect size =0.47) or all students in the two grades (effect size = 0.46).

Two particularly interesting issues are raised by Figure 1. First, in both IAEP2 and TIMSS, the gap in performance between 13-year-olds in the two grades is much larger in such countries as France and Portugal than it is in others (including Ireland). Second, in TIMSS, the difference between grade averages for 13-year-olds seems to be much larger in some countries than the difference between averages for grades when all students are considered. Both findings have a bearing on Ireland's relative performance in IAEP2 and TIMSS, and as we shall see, both issues are interrelated.

In Canada, France, Portugal, Spain and Switzerland, outcomes from IAEP2 and TIMSS suggest strongly that the difference between the performances of 13-year-olds at the two grade levels is much larger than in Ireland. In England and Korea, the differences are smaller. In Hungary, Scotland, Slovenia, and the US, they are about equal to Ireland. Clearly, achievement in some countries is not only affected by the grade the students are in, but also by how old the students are. This is evident when one contrasts the TIMSS

---

large. It should be acknowledged, however, that effect sizes of any magnitude achieve significance only in the context of the circumstances of their interpretation (Durlak, 1995).

results given in Table 1 and 3. It can be seen that at the seventh grade level, while averages for 13-year-olds in Canada, France, Portugal, Spain and Switzerland are below averages for the grade as a whole, the situation is reversed at the eighth grade level. It is also evident when the TIMSS averages for students in the same grade but born in different years are compared (see Figure 2).

Differences between the cohorts (born in different years but in the same grade) are presented in the metric of effect sizes in Figure 2. The data used to calculate the effect sizes are contained in Table 3 (presented earlier) and Table 4.

[Insert Table 4 about here]

It will be recalled that Table 3 contains the TIMSS average scale scores for the 13-year-olds (born in 1981) that were in grades seven and eight. Table 4 contains the TIMSS average scale scores for the younger students in grade 7 i.e. 12-year-olds or students born in 1982 and for the older students in grade 8 i.e. 14-year-olds or students born in 1980.

[Insert Figure 2 about here]

In England, Hungary, Ireland, Korea, Slovenia and the US, year of birth makes little difference to overall achievement for students in the same grade.[6] However, it is clear that year of birth makes a difference in Canada, France, Portugal, Spain and Switzerland. Students born in 1981 (i.e 13-year-olds) in these countries tend to be lower

---

achieving at the seventh grade and higher achieving at the eighth grade than other students at the grade level. Goldstein (1995) noted that grade-based sampling is often complicated by the need to account for policies related to grade repetition and promotion. Contact with educationists in the above countries confirmed that in many instances these issues were indeed contributory factors. For example, in Spain up to 25% of the students in secondary school grades are repeaters (Guillermo Gil, personal communication, January, 1999).[7] An even higher percentage of repeaters are found in grades in Portugal. Discussing the sampling problems in Portugal for IAEP2, Lapointe, Askew and Meade (1992) noted that "the restriction of certain grades in the Portuguese assessment was necessitated by a very dispersed student population resulting from a unique education system that allows students to repeat any grade up to three times" (p. 6). In France, age is considered to be a very good predictor of success because pupils repeat grades when they are not performing well enough to move up. Therefore, the younger students in a grade are higher achieving (Gérard Bonnet, personal communication, Jan 1999). In Canada and Switzerland, the situation is less clear due to the non-centralised nature of the educational system and the fact that policies vary from canton to canton or from province to province.

---

[6] The same holds for Scottish comparisons of students born in 1982 and 1983 at the seventh grade and of students born in 1981 and 1982 at the eighth grade.

[7] Gil wrote (personal communication, Jan 99): In the Spanish educational system we have a strong tradition of grade repetition. Weaker students repeat grades. Up to 25% of the students in a particular grade are repeaters in secondary education. Repeaters are the students with lower achievement and, consequently they are the oldest students in the grade. In all of our research there is a clear difference in achievement between non repeaters and repeaters, that translates into significant differences between student of different ages within a particular grade. 1990 new law of education limits the amount of repetition to two grades for primary and secondary education, but this measure was not in force at the time of TIMSS.

Conclusion

The study described in this paper was prompted by apparent inconsistencies in the relative performances of some countries when the published results of the IAEP2 and TIMSS surveys were compared. However, a more in-depth analysis of the outcomes indicated that, when samples of students tested in both surveys were matched by age and grade, the findings were somewhat more consistent. This was especially true in the case of Ireland. The poor performance of Irish 13-year-olds in IAEP2 resulted mainly from the fact that a majority of 13-year-olds in most other countries were a grade further on in school. The fact is that science achievement in Ireland was not at the low level suggested by the initial IAEP2 results for the 13-year-olds. When these results were broken out by age and grade, Ireland's relative performance in IAEP2 bore a stronger similarity to its performance in TIMSS (around the international average). The age/grade issue also helped to explain some of the inconsistencies in the performance of other countries across the two surveys also, especially at the seventh grade. At the eighth grade and in the case of countries such as France and Switzerland, the presentation of results broken out by age and grade did not provide a satisfactory explanation for the apparent inconsistencies in the initial published results. Other factors such as individual country response rates and coverage of target populations, the overlap between the content tested in the international tests and the content emphasised in curricula, item format (IAEP2 contained multiple-choice items only, TIMSS included multiple-choice, short answer and extended response items), quality control in data collection, and the motivation of students participating in international tests to do well in international tests may also be relevant (see, O'Leary, 1999).

These findings serve to remind us of the inherent dangers in taking the results of international comparative studies at face value. It is always tempting to talk in terms of rank ordering or the "international horse race" because this is the simplest and most straightforward way in which to present country differences. As Mislevy (1995) and Murphy (1996) point out, the rankings of nations enjoys wide popular interest and have immense impact. However, the reality is that ranks have limited meaning at best, and, as we have seen, may even be grossly misleading. Moreover, these findings also make it clear that a reasonable evaluation of country performance often requires an awareness of the context in which students are compared. A weakness of the international IAEP2 report, *Learning Science* (Lapointe, Askew, & Meade, 1992), was that it failed to highlight the fact that grade issues relating to the 13-year-olds tested might be central to an understanding of Ireland's poor performance or Switzerland's excellent performance. Indeed, it seems unusual in the context of considering system variables such as minutes of instruction and class size that grade was not used to explain differences among countries. Even in the Irish report of IAEP2 results, where the percentage of students at each grade level is provided, no mention is made of the fact that it differed considerably from other countries (see Martin, Hickey, & Murchan, 1992).

An interesting finding in this study was that age and grade were shown to have different effects on achievement across different countries. This is an issue that has not received much attention in previous international studies although some attempt was made in the IEA study of reading literacy to adjust country means for age differences (see Appendix E in Elley, 1992). Interestingly, some of the largest effects in this study were associated with Canada, France, Spain and Portugal -- countries where science achievement also seems to differ across age and grade cohorts because of such policies as

grade repetition. Clearly, there may be other countries where such policies also apply and these countries need to be identified and highlighted in the international reports. Above all, an effort should be made to develop procedures that allow for the outcomes of international tests to be adjusted for age, grade and/or policies around grade repetition and social promotion (see Goldstein, 1995). The TIMSS data provide an excellent opportunity to make a beginning on that important work.

At this point in time the popularity of international comparative studies shows no sign of abating. During the 1998/99 school year, the IEA's TIMSS is being administered at the eighth grade in about 40 countries (not including Ireland). Known as TIMSS-Repeat (TIMSS-R), the study will provide participants with information on trends in mathematics and science achievement (see. IEA, 1999). In addition, a new organization has just entered into the arena of international comparative assessments. Beginning in the year 2000, surveys of mathematics, reading, and science literacy will be conducted in over 30 countries every three years under the auspices the Organisation for Economic Co-operation and Development (OECD). This cycle of surveys, known as the Programme for International Student Assessment (PISA), will focus initially on the proficiencies of 15-year-olds. The target population is age-based rather than grade-based because 15 is the highest age at which enrolment in OECD countries is essentially universal. The study aims to measure competencies that are broader and less tied to curricula than has been the case heretofore. Ireland is committed to participation in the first three cycles of data gathering. Given what we now know about the factors that have impinged on performance in past international studies, it seems evident that those same factors should be the focus of very close attention in PISA. Ensuring that policy makers can make useful decisions based on the PISA data demands no less.

References

Beaton, A. E., Martin, M. O., & Mullis, I. V. S. (1997). Providing data for educational policy in an international context: The Third International Mathematics and Science Study (TIMSS). *European Journal of Psychological Assessment, 13*(49-58).

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Cohen, P. (1977). *Statistical power analysis for the behavioural sciences*. New York: Academic Press.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Maentel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Durlak, J. A. (1995). Understanding meta analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 319-352). Washington, DC: American Psychological Association.

ETS. (1992). *IAEP technical report: Volume 1*. Princeton, NJ: Educational Testing Service.

Foy, P. (1998). Sampling issues in international assessments. *International Journal of Educational Research, 29*, 555-568.

Goldstein, H. (1995). *Interpreting international comparisons of student achievement*. Paris: UNESCO.

Goldstein, H. (1997). International comparisons of student achievement. In A. Little & A. Wolf (Eds.), *Assessment in transition: Learning, monitoring and selection in international perspective* (pp. 58-87). Oxford, England: Pergamon.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis 4th edition*. Englewood Cliffs, NJ: Prentice-Hall Inc.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haentszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jaeger, R. M. (1994). Evaluating policy inferences drawn from international comparisons of students' achievement test performances. *Studies in Educational Evaluation, 20*, 23-39.

Keeves, J. P. (1992, April). *Technical issues in the first and second IEA science studies*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Kellaghan, T., & Grisay, A. (1995). International comparisons of student achievement: Problems and prospects. In OECD (Ed.), *Measuring what students learn*. Paris: OECD.

Lapointe, A. E., Askew, J. M., & Mead, N. A. (1992). *Learning science*. Princeton, NJ: Educational Testing Service.

Martin, M. O., Hickey, B. L., & Murchan, D. P. (1992). The second international assessment of educational progress: mathematics and science findings in Ireland. *Irish Journal of Education, 26*, 3-146.

Martin, M. O., & Kelly, D. L. (Eds.). (1996). *Third International Mathematics and Science Study Technical Report Volume 1: Design and development*. Chestnut Hill,

MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Martin, M. O., Mullis, I. V. S., Beaton, A. E., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1997). *Science achievement in the primary school years: IEA's Third International mathematics and Science Study*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

O'Leary, M. (1999). *The validity and consistency of findings from international comparative studies of student achievement: A comparison of outcomes from IAEP2 and TIMSS*. Unpublished doctoral dissertation, Boston College, Massachusetts.

O'Leary, M., Madaus, G., & Kellaghan, T. (1997, July). *The validity and stability of international comparative findings in mathematics and science*. Paper presented at the International Centre for Research on Assessment Conference, London, England.

Plomp, T. (1992). Conceptualizing a comparative educational research framework. *Prospects, 22, 3*, 278-288.

Wiley, D. E., & Wolfe, R. G. (1992). Major survey design issues for the IEA Third International Mathematics and Science Study. *Prospects, 22*, 297-304.

Wolf, F. M. (1986). *Meta Analysis: Qualitative methods for research synthesis*. Newbury Park, CA: Sage.

Table 1

Science Averages of Countries that Participated in IAEP2 and TIMSS (Categorised in Terms of the Significance of Difference of Each Average from the Irish Average)[a]

IAEP2 13-year-olds				TIMSS Grade 7				TIMSS Grade 8			
	$\bar{X}$	se	sd		$\bar{X}$	se	sd		$\bar{X}$	se	sd
Kor	570	2.3	68	Kor	535	2.1	92	Kor	565	1.9	94
Swi	553	3.4	63	Slo	530	2.4	86	Slo	560	2.5	88
Hun	552	2.3	72	Hun	518	3.2	91	**Hun**	**554**	2.8	90
Slo	536	2.2	65	Eng	512	3.5	101	**Eng**	**552**	3.3	106
Can	534	1.5	61	**US**	**508**	5.5	105	**Ire**	**538**	4.5	96
Eng	533	3.9	71	**Can**	**499**	2.3	90	**US**	**534**	4.7	106
Fra	531	2.5	69	**Ire**	**495**	3.5	91	**Can**	**531**	2.6	93
Sco	529	2.8	69	**Swi**	**484**	2.5	82	Swi	522	2.5	91
Spa	525	2.3	61	Spa	477	2.1	80	Sco	517	5.1	100
US	**523**	4.4	68	Sco	468	3.8	94	Spa	517	1.7	78
Ire	**509**	2.5	72	Fra	451	2.6	74	Fra	498	2.5	77
Por	**504**	3.8	72	Por	428	2.1	71	Por	480	2.3	74

[a] Average performance in countries within the shaded area is not statistically significantly different to that in Ireland. Average performance in countries above the shaded area is statistically significantly above that in Ireland. Average performance in countries below the shaded area is statistically significantly below that in Ireland. Statistically significant at the 0.05 level, adjusted for 11comparisons.

Source: International Assessment of Educational Progress (IAEP2), 1991-1992. IEA's Third International Mathematics and Science Study (TIMSS), 1994-1995.

BEST COPY AVAILABLE

Table 2

Distribution of 13-year-olds (born in 1977) Across Grades in the IAEP2 Sample

	Weighted % Grade 7	Weighted % Grade 8	Weighted % Outside Grades 7 and 8
Can	18.9	79.9	1.2 (most above)
Eng	33.3	66.7	0.0
Fra	32.0	56.6	11.4 (most below)
Hun	38.4	57.8	3.7 (most below)
Ire	63.1	35.5	1.4 (all below)
Kor	30.0	67.2	2.8 (all above)
Por	34.3	54.7	11.0 (most below)
Sco	00.5	86.0	13.5 (all above)
Slo	81.3	13.1	5.6 (all below)
Spa	21.0	79.0	0.0
Swi	26.1	71.8	2.1 (all above)
US	38.4	58.1	3.4 (most below)

Source: International Assessment of Educational Progress (IAEP2), 1991-1992.

Table 3

Science Outcomes By Grade in Average Scale Scores for 13-year-olds (born 1977) in IAEP2 and for 13-year-olds (born 1981) in TIMSS (Categorised in Terms of the Significance of the Difference of Each Average from the Irish Average)[a]

	IAEP2 13-year-olds				TIMSS 13-year-olds		
	$\bar{x}$	se	sd		$\bar{x}$	se	sd
				Grade 7			
Kor	561	3.4	69	Kor	537	2.3	90
Hun	541	2.8	67	Slo	534	2.4	85
Slo	537	2.2	62	Hun	521	3.8	92
Swi	527	5.2	64	Eng	519	4.9	103
Eng	**526**	8.3	71	**US**	**503**	6.1	105
US	**506**	6.2	67	**Ire**	**497**	3.8	92
Can	**501**	2.6	64	**Can**	**480**	4.2	91
Ire	**498**	2.7	70	Swi	464	4.9	87
Fra	**494**	2.7	57	Spa	454	4.1	77
Spa	480	3.3	56	Fra	433	3.4	72
Por	469	3.9	57	Por	410	3.1	68
Sco	-	-	-	Sco	-	-	-
				Grade 8			
Kor	574	2.6	67	Slo	572	5.6	85
Hun	566	2.8	69	Kor	566	3.6	95
Slo	565	5.1	62	Hun	561	2.8	85
Swi	563	3.6	59	**Eng**	**548**	3.9	106
Fra	559	2.1	56	**US**	**542**	4.8	103
Can	**541**	1.6	57	**Ire**	**541**	5.0	94
Por	**541**	2.5	54	**Can**	**537**	2.7	92
US	**538**	3.3	62	**Swi**	**532**	2.5	87
Spa	**537**	2.4	57	**Spa**	**526**	2.1	76
Eng	**536**	4.1	70	Sco	519	5.3	100
Ire	**533**	3.6	69	Fra	511	2.8	73
Sco	**527**	3.0	68	Por	494	2.6	73

[a] Average performance in countries within the shaded area is not statistically significantly different to that in Ireland. Average performance in countries above the shaded area is statistically significantly above that in Ireland. Average performance in countries below the shaded area is statistically significantly below that in Ireland. Statistically significant at the 0.05 level, adjusted for 10 comparisons at the seventh grade and 11 comparisons at the eighth grade.

Source: International Assessment of Educational Progress (IAEP2), 1991-1992. IEA's Third International Mathematics and Science Study (TIMSS), 1994-1995.

BEST COPY AVAILABLE

Table 4

Average Scale Score for 12-year-olds (born 1982) in Grade 7 and for 14-year-olds (born 1980) in Grade 8 in TIMSS

	12-year-olds Grade 7			14-year-olds Grade 8		
	$\bar{x}$	se	sd	$\bar{x}$	se	sd
Can	505	2.2	88	510	4.8	92
Eng	509	4.0	100	560	5.4	107
Fra	464	2.7	72	482	4.1	76
Hun	527	3.3	86	556	3.6	91
Ire	500	3.9	87	539	4.8	96
Kor	530	4.0	97	565	2.2	93
Por	439	2.3	71	461	3.3	69
Sco	-	-	-	-	-	-
Slo	536	6.1	87	561	2.6	80
Spa	487	2.3	79	500	3.0	75
Swi	489	2.4	79	500	4.8	93
US	516	6.0	103	530	5.2	107

Source: IEA's Third International Mathematics and Science Study (TIMSS), 1994-1995.

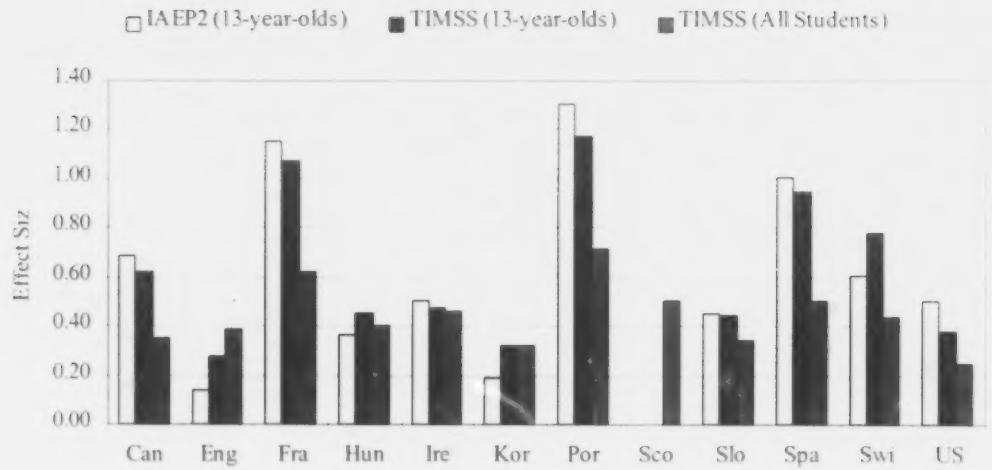

Figure 1. Difference between scale score averages for students in grades 7 and 8 in IAEP2 and TIMSS (expressed in effect sizes).

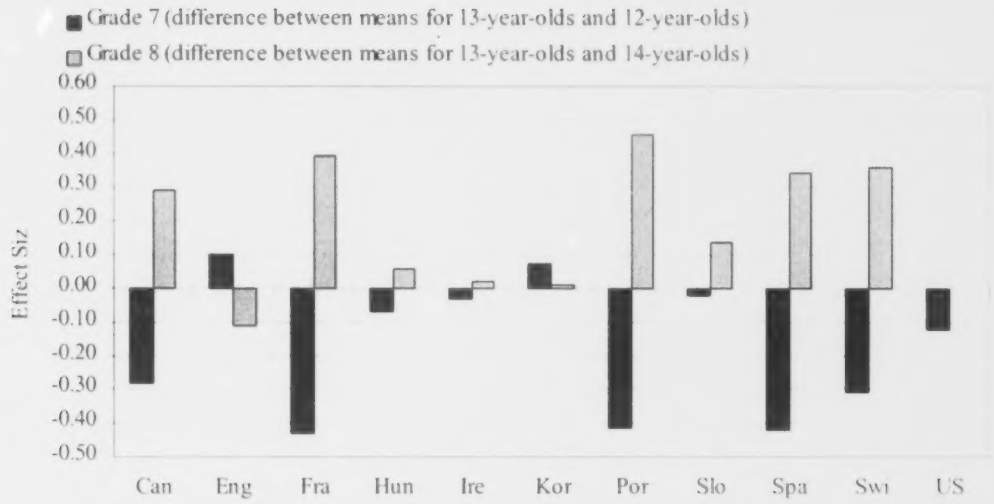

Figure 2. Difference between scale score averages in TIMSS for cohorts of students in the same grade but born in different years (expressed in effect sizes).
Note: Equivalent data for Scotland not available.

Table 1

Average Percents Correct at Grade Eight[a] for 12 Countries Across Different Item Sets in TIMSS (Categorised in Terms of the Significance of Difference of Each Average from the Irish Average)[b]

Overall 135 Items 146 Score Points[c]			Multiple-Choice 102 Items 102 Score Points			Short-Answer 22 items 25 Score Points			Extended-Response 11 items 19 Score Points		
	$\bar{x}$	se		$\bar{x}$	se		$\bar{x}$	se		$\bar{x}$	se
Kor	65.5	0.3	Kor	70.2	0.4	Kor	62.1	0.9	**Eng**	**54.6**	0.9
Slo	61.7	0.5	Slo	66.5	0.5	Eng	61.9	1.0	**Ire**	**52.8**	1.2
Eng	**61.3**	0.6	Hun	65.6	0.5	**Hun**	**59.0**	1.1	**Kor**	**52.6**	0.7
Hun	**60.7**	0.6	**Eng**	**63.7**	0.6	**Slo**	**58.0**	0.9	Can	48.7	0.7
Can	**58.7**	0.5	**Can**	**62.2**	0.5	**Can**	**57.3**	0.6	Swi	48.4	0.8
Ire	**58.4**	0.9	**US**	**61.9**	0.9	**Spa**	**56.0**	0.8	Slo	47.7	1.1
US	**58.3**	1.0	**Ire**	**61.3**	0.9	**Ire**	**55.7**	1.2	Sco	47.6	1.2
Swi	**56.3**	0.5	**Swi**	**59.8**	0.5	**US**	**54.5**	1.2	US	47.1	1.3
Spa	**55.6**	0.4	**Spa**	**59.2**	0.4	**Swi**	**52.7**	0.7	Hun	43.6	1.0
Sco	**55.3**	1.0	**Sco**	**58.7**	1.0	**Sco**	**52.4**	1.3	Spa	41.8	0.6
Fra	53.7	0.6	Fra	57.9	0.6	Fra	49.9	1.0	Fra	40.7	0.9
Por	49.9	0.6	Por	55.5	0.6	Por	43.7	0.9	Por	34.1	0.7
Int'l[d]	58.0			61.9			55.3			46.6	

[a] Grade 8 in most countries.

[b] Average performance in countries within the shaded area is not statistically significantly different to that in Ireland. Average performance in countries above the shaded area is statistically significantly above that in Ireland. Average performance in countries below the shaded area is statistically significantly below that in Ireland. Statistically significant at the 0.05 level, adjusted for 11 comparisons.

[c] Some of the TIMSS science items had more than one part and this resulted in a total of 146 score points in all.

[d] The average of the 12 country averages.

Source: IEA's Third International Mathematics and Science Study (TIMSS), 1994-1995.

BEST COPY AVAILABLE

Table 2

Percentages Omitting Individual Multiple-Choice Science Items in TIMSS (Lower Grade)

Item ID	Can	Eng	Fra	Hun	Ire	Kor	Por	Sco	Slo	Spa	Swi	US
E11	1	1	6	8	2	0	5	3	6	4	7	1
G10	1	1	9	3	2	1	7	1	1	2	4	1
I11	3	9	5	10	3	0	9	7	7	13	3	3
K11	1	3	5	6	4	0	2	3	5	5	6	1
K14	4	7	3	6	2	0	7	1	4	6	5	3
L01	1	3	8	4	4	0	8	3	4	6	5	1
N01	2	2	1	5	2	1	4	2	3	6	6	2
N09	2	5	11	9	1	0	7	3	3	7	5	1
O11	1	6	9	7	1	0	7	1	3	3	7	1
O15	1	1	10	1	1	0	7	0	1	3	5	3
Q13	1	1	4	5	2	1	8	1	3	7	4	2

Source: TIMSS (1996)

Table 3

Percentages at the Eighth Grade Omitting Extended Response Science Items in TIMSS

Item ID	Can	Eng	Fra	Hun	Ire	Kor	Por	Sco	Slo	Spa	Swi	US
L04	20.1	7.9	10.3	11.2	7.3	7.6	10.2	11.1	13.7	13.0	7.5	8.5
M11	3.9	1.7	4.3	8.0	3.5	1.3	12.2	5.2	3.4	9.5	5.3	10.4
O14	1.1	1.3	6.1	4.7	3.4	0.4	2.9	0.6	1.4	3.7	5.8	0.9
W01 A	2.2	1.8	10.3	12.1	2.5	3.2	5.4	2.6	3.0	6.6	7.9	1.7
B	6.9	2.2	20.6	9.5	3.2	5.7	19.0	6.9	10.6	15.8	7.7	2.6
W02	5.9	9.8	11.2	15.8	6.5	18.8	16.6	9.5	7.1	13.3	7.1	7.8
X01	18.6	14.1	31.3	39.6	25.2	23.2	29.3	27.0	20.6	34.5	14.7	14.1
X02 A	3.5	3.9	9.9	5.1	3.8	4.1	7.1	5.5	5.6	4.5	4.9	3.4
B	2.3	4.5	8.8	8.9	4.0	2.9	6.3	3.2	1.4	4.4	4.9	2.0
Y01	2.2	2.4	7.0	18.0	3.8	5.8	5.9	3.6	2.6	8.7	6.7	2.2
Y02	7.4	5.1	12.9	9.6	3.2	6.5	14.9	7.7	15.3	10.5	5.2	8.2
Z01 A	6.7	3.5	16.4	14.1	4.6	4.3	9.2	4.3	14.0	11.6	10.3	6.2
B	10.8	7.0	16.3	-	6.2	2.8	0.6	18.5	15.6	17.5	10.4	6.3
C	28.0	24.3	45.4	-	20.7	6.0	42.0	33.0	33.3	35.9	39.0	26.8
Z02 A	1.6	0.5	3.0	12.9	0.5	6.2	4.4	2.2	9.8	1.7	1.5	1.0
B	11.9	16.5	22.2	16.4	12.5	9.4	37.1	19.7	19.6	14.4	17.2	13.1

Source: TIMSS (1996)

Table 4

Percentages at the Eighth Grade Not Reaching the Final Science Items in the TIMSS Booklets

Item ID	Can	Eng	Fra	Hun	Ire	Kor	Por	Sco	Slo	Spa	Swi	US
W02	4.7	1.9	11.6	10.6	30.	6.4	14.8	3.6	5.2	12.2	6.5	3.0
X02A	4.1	2.9	8.4	12.5	4.4	6.0	8.7	7.2	6.1	9.9	5.1	4.5
XO2B	5.8	4.1	16.1	15.7	6.1	7.3	14.1	10.4	9.5	12.5	8.2	6.9
Y02	1.9	0.2	3.4	8.6	1.9	3.2	4.0	2.2	1.6	5.9	3.2	1.7
Z02A	5.5	4.1	16.8	15.0	5.9	3.5	12.9	6.3	17.1	8.8	8.0	5.8
Z02B	6.9	4.6	19.8	20.6	6.4	9.7	17.0	8.5	25.6	10.5	9.5	6.8

Source: TIMSS (1996)

Table 5

Classification of the Extended-Response Items in TIMSS by Content Category and Performance Expectation

Item ID	Content Category	Performance Expectation
L04	Physics	Applying and Investigating Scientific Principles
M11	Life Science	Understanding Complex Information
O14	Earth Science	Applying and Investigating Scientific Principles
W01	Earth Science	Applying and Investigating Scientific Principles
W02	Earth Science	Applying and Investigating Scientific Principles
X01	Life Science	Applying and Investigating Scientific Principles
X02	Life Science	Applying and Investigating Scientific Principles
Y01	Physics	Applying and Investigating Scientific Principles
Y02	Physics	Applying and Investigating Scientific Principles
Z01	Chemistry	Applying and Investigating Scientific Principles
Z02	Environmental Issues	Applying and Investigating Scientific Principles

Source: TIMSS (1996)

Table 6
The Test-Curriculum Match for Extended-Response Items in TIMSS

Item ID	Can	Eng	Fra	Hun	Ire	Kor	Por	Sco	Slo	Spa	Swi	US
L04	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
M11	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
O14	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes
W01	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
W02	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
X01	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
X02	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Yes
Y01	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	No	Yes
Y02	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	No	Yes
Z01	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes
Z02	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Total Yes	11	11	5	11	6	3	11	10	11	11	8	11

Source: TIMSS (1996)